(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 784 311 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
16.07.1997 Bulletin 1997/29

(51) Int. Cl.[6]: **G10L 3/00**

(21) Application number: **96118504.8**

(22) Date of filing: **19.11.1996**

(84) Designated Contracting States:
CH DE FR GB IT LI NL SE

(30) Priority: **12.12.1995 FI 955947**

(71) Applicant: **NOKIA MOBILE PHONES LTD.**
24101 Salo (FI)

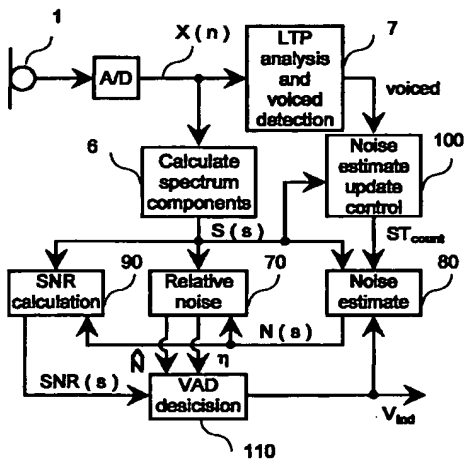
(72) Inventors:
- **Vähätalo, Antti**
  33610 Tampere (FI)
- **Paajanen, Erkki**
  33540 Tampere (FI)
- **Häkkinen, Juha**
  33710 Tampere (FI)

(74) Representative: **Johansson, Folke Anders**
Nokia Mobile Phones Ltd.,
P.O. Box 100
00045 Nokia Group (FI)

(54) Method and device for voice activity detection and a communication device

(57) The invention concerns a voice activity detection device in which an input speech signal (x(n)) is divided in subsignals (S(s)) representing specific frequency bands and noise (N(s)) is estimated in the subsignals. On basis of the estimated noise in the subsignals, subdecision signals (SNR(s)) are generated and a voice activity decision (V_{ind}) for the input speech signal is formed on basis of the subdecision signals. Spectrum components of the input speech signal and a noise estimate are calculated and compared. More specifically a signal-to-noise ratio is calculated for each subsignal and each signal-to-noise ratio represents a subdecision signal (SNR(s)). From the signal-to-noise ratios a value proportional to their sum is calculated and compared with a threshold value and a voice activity decision signal (V_{ind}) for the input speech signal is formed on basis of the comparison.

Fig.2



EP 0 784 311 A1

Description

This invention relates to a voice activity detection device comprising means for detecting voice activity in an input signal, and for making a voice activity decision on basis of the detection. Likewise the invention relates to a method for detecting voice activity and to a communication device including voice activity detection means.

A Voice Activity Detector (VAD) determines whether an input signal contains speech or background noise. A typical application for a VAD is in wireless communication systems, in which the voice activity detection can be used for controlling a discontinuous transmission system, where transmission is inhibited when speech is not detected. A VAD can also be used in e.g. echo cancellation and noise cancellation.

Various methods for voice activity detection are known in prior art. The main problem is to reliably detect speech from background noise in noisy environments. Patent publication US 5,459,814 presents a method for voice activity detection in which an average signal level and zero crossings are calculated for the speech signal. The solution achieves a method which is computationally simple, but which has the drawback that the detection result is not very reliable. Patent publications WO 95/08170 and US 5,276,765 present a voice activity detection method in which a spectral difference between the speech signal and a noise estimate is calculated using LPC (Liner Prediction Coding) parameters. These publications also present an auxiliary VAD detector which controls updating of the noise estimate. The VAD methods of all the above mentioned publications have problems to reliably detect speech when speech power is low compared to noise power.

The present invention concerns a voice activity detection device in which an input speech signal is divided in subsignals representing specific frequency bands and voice activity is detected in the subsignals. On basis of the detection of the subsignals, subdecision signals are generated and a voice activity decision for the input speech signal is formed on basis of the subdecision signals. In the invention spectrum components of the input speech signal and a noise estimate are calculated and compared. More specifically a signal-to-noise ratio is calculated for each subsignal and each signal-to-noise ratio represents a subdecision signal. From the signal-to-noise ratios a value proportional to their sum is calculated and compared with a threshold value and a voice activity decision signal for the input speech signal is formed on basis of the comparison.

For obtaining the signal-to-noise ratios for each subsignal a noise estimate is calculated for each subfrequency band (i.e. for each subsignal). This means that noise can be estimated more accurately and the noise estimate can also be updated separately for each subfrequency band. A more accurate noise estimate will lead to a more accurate and reliable voice activity detection decision. Noise estimate accuracy is also improved by using the speech/noise decision of the voice activity detection device to control the updating of the background noise estimate.

A voice activity detection device and a communication device according to the invention is characterized by that it comprises means for dividing said input signal in subsignals representing specific frequency bands, means for estimating noise in the subsignals, means for calculating subdecision signals on basis of the noise in the subsignals, and means for making a voice activity decision for the input signal on basis of the subdecision signals.

A method according to the invention is characterized by that it comprises the steps of dividing said input signal in subsignals representing specific frequency bands, estimating noise in the subsignals, calculating subdecision signals on basis of the noise in the subsignals, and making a voice activity decision for the input signal on basis of the subdecision signals.

In the following, the invention is illustrated in more detail, referring to the enclosed figures, in which

- fig. 1 presents a block diagram of a surroundings of use of a VAD according to the invention,
- fig. 2 presents in the form of a block diagram a realization of a VAD according to the invention,
- fig. 3 presents a realization of the power spectrum calculation block in fig. 2,
- fig. 4 presents an alternative realization of the power spectrum calculation block,
- fig. 5 presents in the form of a block diagram another embodiment of the device according to the invention,
- fig. 6 presents in the form of a block diagram a realization of a windowing block,
- fig. 7 presents subsequent speech signal frames in windowing according to the invention,
- fig. 8 presents a realization of a squaring block,
- fig. 9 presents a realization of a spectral recombination block,
- fig. 10 presents a realization of a block for calculation of relative noise level,
- fig. 11 presents an arrangement for calculating a background noise model,
- fig. 12 presents in form of a block diagram a realization of a VAD decision block, and
- fig. 13 presents a mobile station according to the invention.

Figure 1 shows shortly the surroundings of use of the voice activity detection device 4 according to the invention. The parameter values presented in the following description are exemplary values and describe one embodiment of the invention, but they do not by any means limit the function of the method according to the invention to only certain parameter values. Referring to figure 1 a signal coming from a microphone 1 is sampled in an A/D converter 2. As exemplary

values it is assumed that the sample rate of the A/D converter 2 is 8000 Hz, the frame length of the speech codec 3 is 80 samples, and each speech frame comprises 10 ms of speech. The VAD device 4 can use the same input frame length as the speech codec 3 or the length can be an even quotient of the frame length used by the speech codec. The coded speech signal is fed further in a transmission branch, e.g. to a discontinous transmission handler 5, which controls transmission according to a decision V_{ind} received from the VAD 4.

One embodiment of the voice activity detection device according to the invention is described in more detail in figure 2. A speech signal coming from the microphone 1 is sampled in an A/D-converter 2 into a digital signal x(n). An input frame for the VAD device in Fig. 2 is formed by taking samples from digital signal x(n). This frame is fed into block 6, in which power spectrum components presenting power in predefined bands are calculated. Components proportional to amplitude or power spectrum of the input frame can be calculated using an FFT, a filter bank, or using linear predictor coefficients. This will be explained in more detail later. If the VAD operates with a speech codec that calculates linear prediction coefficients then those coefficients can be received from the speech codec.

Power spectrum components P(f) are calculated from the input frame using first Fast Fourier Transform (FFT) as presented in figure 3. In the example solution it is assumed that the length of the FFT calculation is 128. Additionally, power spectrum components P(f) are recombined to calculation spectrum components S(s) reducing the number of spectrum components from 65 to 8.

Referring to Fig. 3 a speech frame is brought to windowing block 10, in which it is multiplied by a predetermined window. The purpose of windowing is in general to enhance the quality of the spectral estimate of a signal and to divide the signal into frames in time domain. Because in the windowing used in this example windows partly overlap, the overlapping samples are stored in a memory (block 15) for the next frame. 80 samples are taken from the signal and they are combined with 16 samples stored during the previous frame, resulting in a total of 96 samples. Respectively out of the last collected 80 samples, the last 16 samples are stored for being used in calculating the next frame.

The 96 samples given this way are multiplied in windowing block 10 by a window comprising 96 sample values, the 8 first values of the window forming the ascending strip I_U of the window, and the 8 last values forming the descending strip I_D of the window, as presented in figure 7. The window I(n) can be defined as follows and is realized in block 11 (figure 6):

$$I(n)= (n+1)/9 = I_U \qquad n=0,..,7 \tag{1}$$

$$I(n)=1 = I_M \qquad n=8,..,87$$

$$I(n)=(96-n)/9 = I_D \qquad n=88,..,95$$

Realizing of windowing (block 11) digitally is prior known to a person skilled in the art of digital signal processing. It has to be notified that in the window the middle 80 values (n=8,..87 or the middle strip I_M) are equal to 1, and accordingly multiplication by them does not change the result and the multiplication can be omitted. Thus only the first 8 samples and the last 8 samples in the window need to be multiplied. Because the length of an FFT has to be a power of two, in block 12 (figure 6) 32 zeroes (0) are added at the end of the 96 samples obtained from block 11, resulting in a speech frame comprising 128 samples. Adding samples at the end of a sequence of samples is a simple operation and the realization of block 12 digitally is within the skills of a person skilled in the art.

After windowing has been carried out in windowing block 10, the spectrum of a speech frame is calculated in block 20 employing the Fast Fourier Transform, FFT. Samples x(0),x(1),..,x(n); n=127 (or said 128 samples) in the frame arriving to FFT block 20 are transformed to frequency domain employing real FFT (Fast Fourier Transform), giving frequency domain samples X(0),X(1),..,X(f);f=64 (more generally f=(n+1)/2) , in which each sample comprises a real component $X_r(f)$ and an imaginary component $X_i(f)$:

$$X(f) = X_r(f) + jX_i(f), \; f=0,..,64 \tag{2}$$

Realizing Fast Fourier Transform digitally is prior known to a person skilled in the art. The real and imaginary components obtained from the FFT are squared and added together in pairs in squaring block 50, the output of which is the power spectrum of the speech frame. If the FFT length is 128, the number of power spectrum components obtained is 65, which is obtained by dividing the length of the FFT transformation by two and incrementing the result with 1, in other words the length of FFT/2 + 1 . Accordingly, the power spectrum is obtained from squaring block 50 by calculating the sum of the second powers of the real and imaginary components, component by component:

$$P(f)=X_r^2(f)+X_i^2(f), \; f=0,..,64 \tag{3}$$

The function of squaring block 50 can be realized, as is presented in figure 8, by taking the real and imaginary components to squaring blocks 51 and 52 (which carry out a simple mathematical squaring, which is prior known to be car-

ried out digitally) and by summing the squared components in a summing unit 53. In this way, as the output of squaring block 50, power spectrum components P(0), P(1),..,P(f);f=64 are obtained and they correspond to the powers of the components in the time domain signal at different frequencies as follows (presuming that 8 kHz sampling frequency is used):

$P(f)$ for values f = 0,...,64 corresponds to middle frequencies (f · 4000/64 Hz) (4)

After this 8 new power spectrum components, or power spectrum component combinations S(s), s =0,..7 are formed in block 60 and they are here called calculation spectrum components. The calculation spectrum components S(s) are formed by summing always 7 adjacent power spectrum components P(f) for each calculation spectrum component S(s) as follows:

$$S(0)= P(1)+P(2)+..+P(7) \quad (5)$$

$$S(1)= P(8)+P(9)+..+P(14)$$

$$S(2)= P(15)+P(16)+..+P(21)$$

$$S(3)= P(22)+..+P(28)$$

$$S(4)= P(29)+..+P(35)$$

$$S(5)= P(36)+..+P(42)$$

$$S(6)= P(43)+..+P(49)$$

$$S(7)= P(50)+..+P(56)$$

This can be realized, as presented in figure 9, utilizing counter 61 and summing unit 62, so that the counter 61 always counts up to seven and, controlled by the counter, summing unit 62 always sums seven subsequent components and produces a sum as an output. In this case the lowest combination component S(0) corresponds to middle frequencies [62.5 Hz to 437.5 Hz] and the highest combination component S(7) corresponds to middle frequencies [3125 Hz to 3500 Hz]. The frequencies lower than this (below 62.5 Hz) or higher than this (above 3500 Hz) are not essential for speech and can be ignored.

Instead of using the solution of Figure 3, power spectrum components P(f) can also be calculated from the input frame using a filter bank as presented in figure 4. The filter bank comprises bandpass filters $H_j(z)$, j=0,...,7; covering the frequency band of interest. The filter bank can be either uniform or composed of variable bandwidth filters. Typically, the filter bank outputs are decimated to improve efficiency. The design and digital implementation of filter banks is known to a person skilled in the art. Sub-band samples $z_j(i)$ in each band j are calculated from the input signal x(n) using filter $H_j(z)$. Signal power at each band can be calculated as follows:

$$S(j) = \sum_{i=0}^{L-1} z_j(i) \cdot z_j(i) \quad (6)$$

where, L is the number of samples in the sub-band within one input frame.

When a VAD is used with a speech codec, the calculation spectrum components S(s) can be calculated using Linear Prediction Coefficients (LPC), which are calculated by most of the speech codecs used in digital mobile phone systems. Such an arrangement is presented in figure 5. LPC coefficients are calculated in a speech codec 3 using a technique called linear prediction, where a linear filter is formed. The LPC coefficients of the filter are direct order coefficients d(i), which can be calculated from autocorrelation coefficients ACF(k). As will be shown below, the direct order coefficients d(i) can be used for calculating calculation spectrum components S(s). The autocorrelation coefficients ACF(k), which can be calculated from input frame samples x(n), can be used for calculating the LPC coefficients. If LPC coefficients or ACF(k) coefficients are not available from the speech codec, they can be calculated from the input frame.

Autocorrelation coefficients ACF(k) are calculated in the speech codec 3 as follows:

$$ACF(k) = \sum_{i=k}^{N} x(i)x(i-k),\ k=0,1,..,M \tag{7}$$

where,

N is the number of samples in the input frame,
M is the LPC order (e.g., 8), and
x(i) are the samples in the input frame.

LPC coefficients d(i), which present the impulse response of the short term analysis filter, can be calculated from the autocorrelation coefficients ACF(k) using a previously known method, e.g., the Schur recursion algorithm or the Levinson-Durbin algorithm.

Amplitude at desired frequency is calculated in block 8 shown in figure 5 from the LPC values using Fast Fourier Transform (FFT) according to following equation:

$$A(k) = \frac{1}{M^{1/2}}\left|\sum_{i=0}^{M-1} d(i)e^{-ij2\pi k/K}\right| \tag{8}$$

where,

K is a constant, e.g. 8000
k corresponds to a frequency for which power is calculated (i.e., A(k) corresponds to frequency k/K*fs , where fs is the sample frequency), and
M is the order of the short term analysis.

The amplitude of a desired frequency band can be estimated as follows

$$A(k1,k2) = \frac{1}{M^{1/2}}\left|\sum_{i=0}^{M-1} d(i)C(k1,k2,i)\right| \tag{9}$$

where

k1 is the start index of the frequency band and k2 is the end index of the frequency band.

The coefficients $C(k1,k2,i)$ can be calculated forehand and they can be saved in a memory (not shown) to reduce the required computation load. These coefficients can be calculated as follows:

$$C(k1,k2,i) = \sum_{k=k1}^{k2} e^{-ij2\pi k/K} \tag{10}$$

An approximation of the signal power at calculation spectrum component S(s) can be calculated by inverting the square of the amplitude A(k1,k2) and by multiplying with ACF(0). The inversion is needed because the linear predictor coefficients presents inverse spectrum of the input signal. ACF(0) presents signal power and it is calculated in the equation 7.

$$S(s) = \frac{ACF(0)}{A(k1,k2)^2} \tag{11}$$

where each calculation spectrum component S(s) is calculated using specific constants k1 and k2 which define the band limits.
Above different ways of calculating the power (calculation) spectrum components S(s) have been described.

Further in Fig. 2 the spectrum of noise N(s), s=0,..,7 is estimated in estimation block 80 (presented in more detail in figure 11) when the voice activity detector does not detect speech. Estimation is carried out in block 80 by calculating recursively a time-averaged mean value for each spectrum component S(s), s=0,..,7 of the signal brought from block 6:

$$N_n(s)=\lambda(s)N_{n-1}(s)+(1-\lambda(s))S(s) \qquad s = 0,...,7. \tag{12}$$

In this context $N_{n-1}(s)$ means a calculated noise spectrum estimate for the previous frame, obtained from memory 83, as presented in figure 11, and $N_n(s)$ means an estimate for the present frame (n = frame order number) according to the equation above. This calculation is carried out preferably digitally in block 81, the inputs of which are the spectrum components S(s) from block 6, the estimate for the previous frame $N_{n-1}(s)$ obtained from memory 83 and the value for time-constant variable λ(s) calculated in block 82. The updating can be done using faster time-constant when input spectrum components are S(s) lower than noise estimate $N_{n-1}(s)$ components. The value of the variable λ(s) is determined according to the next table (typical values for λ(s)):

$S(s) < N_{n-1}(s)$	(V_{ind}, ST_{count})	λ(s)
Yes	(0,0)	0.85
No	(0,0)	0.9
Yes	(0,1)	0.85
No	(0,1)	0.9
Yes	(1,0)	0.9
No	(1,0)	1 (no updating)
Yes	(1,1)	0.9
No	(1,1)	0.95

The values V_{ind} and ST_{count} are explained more closely later on.

In following the symbol $N(s)$ is used for the noise spectrum estimate calculated for the present frame. The calculation according to the above estimation is preferably carried out digitally. Carrying out multiplications, additions and subtractions according to the above equation digitally is well known to a person skilled in the art.

Further in Fig. 2 a ratio SNR(s), s=0,..,7 is calculated from input spectrum $S(s)$ and noise spectrum $N(s)$, component by component, in calculation block 90 and the ratio is called signal-to-noise ratio:

$$SNR(s) = \frac{S(s)}{N(s)}. \tag{13}$$

The signal-to-noise ratios SNR(s) represent a kind of voice activity decisions for each frequency band of the calculation spectrum components. From the signal-to-noise ratios SNR(s) it can be determined whether the frequency band signal contains speech or noise and accordingly it indicates voice activity. The calculation block 90 is also preferably realized digitally, and it carries out the above division. Carrying out a division digitally is as such prior known to a person skilled in the art.

In Fig. 2 relative noise level is calculated in block 70, which is more closely presented in figure 10, and in which the time averaged mean value for speech $\hat{S}(n)$ is calculated using the power spectrum estimate $S(s)$, S=0,..,7. The time averaged mean value $\hat{S}(n)$ is updated when speech is detected. First the mean value $\bar{S}(n)$ of power spectrum components in the present frame is calculated in block 71, into which spectrum components S(s) are obtained as an input from block 60, as follows:

$$\bar{S}(n) = \frac{1}{8}\sum_{s=0}^{7} S(s). \tag{14}$$

The time averaged mean value $\hat{S}(n)$ is obtained by calculating in block 72 (e.g., recursively) based upon a time averaged mean value $\hat{S}(n-1)$for the previous frame, which is obtained from memory 78, in which the calculated time

averaged mean value has been stored during the previous frame, the calculation spectrum mean value $\bar{S}(n)$ obtained from block 71, and time constant α which has been stored in advance in memory 79a:

$$\hat{S}(n) = \alpha\ \hat{S}(n\text{-}1)+(1\text{-}\alpha)\bar{S}(n), \tag{15}$$

in which n is the order number of a frame and α is said time constant, the value of which is from 0.0 to 1.0, typically between 0.9 to 1.0. In order not to contain very weak speech in the time averaged mean value (e.g. at the end of a sentence), it is updated only if the mean value of the spectrum components for the present frame exceeds a threshold value dependent on time averaged mean value. This threshold value is typically one quarter of the time averaged mean value. The calculation of the two previous equations is preferably executed digitally.

Correspondingly, the time averaged mean value of noise power $\hat{N}(_n)$ is obtained from calculation block 73 by using the power spectrum estimate of noise $N(s)$, s=0,..,7 and component mean value $\bar{N}(n)$ calculated from it according to the next equation:

$$\hat{N}(_n) = \beta\ \hat{N}(n\text{-}1)+(1\text{-}\beta)\bar{N}(n), \tag{16}$$

in which β is a time constant, the value of which is 0.0. to 1.0, typically between 0.9 to 1.0. The noise power time averaged mean value is updated in each frame. The mean value of the noise spectrum components $\bar{N}(n)$ is calculated in block 76, based upon spectrum components N(s), as follows:

$$\bar{N}(n)=\frac{1}{8}\sum_{s=0}^{7} N(s) \tag{17}$$

and the noise power time averaged mean value $\hat{N}(n - 1)$ for the previous frame is obtained from memory 74, in which it was stored during the previous frame. The relative noise level η is calculated in block 75 as a scaled and maximum limited quotient of the time averaged mean values of noise and speech

$$\eta = \min\left(max\_\eta, \kappa\frac{\bar{N}}{\bar{S}}\right), \tag{18}$$

in which κ is a scaling constant (typical value 4.0), which has been stored in advance in memory 77, and max_n is the maximum value of relative noise level (typically 1.0), which has been stored in memory 79b.

For producing a VAD decision in the device in Fig. 2, a distance D_{SNR} between input signal and noise model is calculated in the VAD decision block 110 utilizing signal-to-noise ratio SNR(s) , which by digital calculation realizes the following equation:

$$D_{SNR} = \sum_{s=s\_l}^{s\_h} \upsilon_s SNR(s); \tag{19}$$

in which s_l and s_h are the index values of the lowest and highest frequency components included and v_s = component weighting coefficient, which are predetermined and stored in advance in a memory, from which they are retrieved for calculation. Typically, all signal-to-noise estimate value components are used ($s\_l=0$ and $s\_h=7$), and they are weighted equally: v_s = 1.0/8.0; s=0,..,7.

The following is a closer description of the embodiment of a VAD decision block 110, with reference to figure 12. A summing unit 111 in the voice activity detector sums the values of the signal-to-noise ratios SNR(s), obtained from different frequency bands, whereby the parameter D_{SNR}, describing the spectrum distance between input signal and noise model, is obtained according to the above equation (19), and the value D_{SNR} from the summing unit 111 is compared with a predetermined threshold value vth in comparator unit 112. If the threshold value vth is exceeded, the frame is regarded to contain speech. The summing can also be weighted in such a way that more weight is given to the frequencies, at which the signal-to-noise ratio can be expected to be good. The output and decision of the voice activity detector can be presented with a variable V_{ind}, for the values of which the following conditions are obtained:

$$\begin{cases} V_{ind} = 1; & D_{SNR} > vth \\ V_{ind} = 0; & D_{SNR} \le vth \end{cases} \quad (20)$$

Because the VAD controls the updating of background spectrum estimate *N(s)*, and the latter on its behalf affects the function of the voice activity detector in a way described above, it is possible that both noise and speech is indicated as speech ($V_{ind=1}$) if the background noise level suddenly increases. This further inhibits update of the background spectrum estimate *N(s)*. To prevent this, the time (number of frames) during which subsequent frames are regarded not to contain speech is monitored. Subsequent frames, which are stationary and are not indicated voiced are assumed not to contain speech.

In block 7 in figure 2, Long Term Prediction (LTP) analysis, which is also called pitch analysis, is calculated. Voiced detection is done using long term predictor parameters. The long term predictor parameters are the lag (i.e. pitch period) and the long term predictor gain. Those parameters are calculated in most of the speech coders. Thus if a voice activity detector is used besides a speech codec (as described in Fig. 5), those parameters can be obtained from the speech codec.

The long term prediction analysis can be calculated from an amount of samples M which equals frame length N, or the input frame length can be divided to sub-frames (e.g. 4 sub-frames, 4*M=N) and long term parameters are calculated separately from each sub-frame. The division of the input frame into these sub-frames is done in the LTP analysis block 7 (Fig. 2). The sub-frame samples are denoted xs(i).

Accordingly, in block 7 first auto-correlation R(l) from the sub-frame samples xs(i) is calculated,

$$R(l) = \sum_{i=0}^{M} xs(i) \cdot xs(i-l) \quad (21)$$

where

l=Lmin,...,Lmax (e.g. Lmin=40, Lmax=160)

Last Lmax samples from the old sub-frames must be saved for the above mentioned calculation.

Then a maximum value Rmax from the R(l) is searched so that Rmax=max(R(l)), where l=40,...,160.

The long term predictor lag LTP_lag(j) is the index l with corresponds to Rmax. Variable j indicates the index of the sub-frame (j=0..3).

LTP_gain can be calculated as follows:

LTP_gain(j)=Rmax/Rtot

where

$$Rtot = \sum_{i=0}^{N} xs(i - LTP\_lag(j))^2 \quad (22)$$

A parameter presenting the long term predictor lag gain of a frame (LTP_gain_sum) can be calculated by summing the long term predictor lag gains of the sub-frames (LTP_gain(j))

$$LTP\_gain\_sum = \sum_{j=0}^{3} LTP\_gain(j) \quad (23)$$

If the LTP_gain_sum is higher than a fixed threshold thr_lag, the frame is indicated to be voiced:

```
If (LTP_gain_sum > thr_lag)
   voiced = 1
else
```

voiced = 0

Further in Fig 2 an average noise spectrum estimate NA(s) is calculated in block 100 as follows:

$$NA_n(s) = aNA_{n-1}(s)+(1-a)S(s) \qquad s = 0,...,7 \tag{24}$$

where a is a time constant of value 0<a<1 (e.g. 0,9).

Also a spectrum distance D between the average noise spectrum estimate NA(s) and the spectrum estimate S(s) is calculated in block 100 as follows:

$$D = \sum_{s=0}^{7} \frac{\max(NA(s),S(s))}{\min(NA(s),S(s),Low\_Limit)} \tag{25}$$

Low_Limit is a small constant, which is used to keep the division result small when the noise spectrum or the signal spectrum at some frequency band is low.

If the spectrum distance D is larger than a predetermined threshold Dlim, a stationarity counter stat_cnt is set to zero. If the spectrum distance D is smaller that the threshold Dlim and the signal is not detected voiced (voiced = 0), the stationarity counter is incremented. The following conditions are received for the stationarity counter:

If (D > Dlim)
  stat_cnt = 0
if (D<Dlim and voiced =0)
  stat_cnt = stat_cnt+1

Block 100 gives an output stat_cnt which is reset to zero when V_{ind} gets a value 0 to meet the following condition:

if (V_{ind} = 0)
stat_cnt =0

If this number of subsequent frames exceeds a predetermined threshold value *max_spf*, the value of which is e.g. 50, the value of ST_{COUNT} is set at 1. This provides the following conditions for an output ST_{COUNT} in relation to the counter value stat_cnt:

If (stat_cnt > *max_spf*)
  ST_{COUNT} = 1
else
  ST_{COUNT} = 0

Additionally, in the invention the accuracy of background spectrum estimate *N(s)* is enhanced by adjusting said threshold value *vth* of the voice activity detector utilizing relative noise level η (which is calculated in block 70). In an environment in which the signal-to-noise ratio is very good (or the relative noise level η is low), the value of the threshold *vth* is increased based upon the relative noise level η. Hereby interpreting rapid changes in background noise as speech is reduced. Adaptation of the threshold value *vth* is carried out in block 113 according to the following:

$$vth1 = \max(vth\_min1,\ vth\_fix1 - vth\_slope1 \cdot \eta), \tag{26}$$

in which *vth_fix1*, vth_min1, and *vth_slope1* are positive constants, typical values for which are e.g.: *vth_fix1*=2.5; vth_min1=2.0; *vth_slope1*=8.0.

In an environment with a high noise level, the threshold is decreased to decrease the probability that speech is detected as noise. The mean value of the noise spectrum components $\hat{N}(_n)$ is then used to decrease the threshold *vth* as follows

$$vth2 = \min(vth1, vth\_fix2 - vth\_slope2 \cdot \hat{N}(n)) \tag{27}$$

in which vth_fix2 and vth_slope2 are positive constants. Thus if the mean value of the noise spectrum components $\hat{N}(_n)$ is large enough, the threshold vht2 is lower that the theshold vth1.

The voice activity detector according to the invention can also be enhanced in such a way that the threshold *vth2* is further decreased during speech bursts. This enhances the operation, because as speech is slowly becoming more quiet it could happen otherwise that the end of speech will be taken for noise. The additional threshold adaptation can be implemented in the following way (in block 113):

First, D_{SNR} is limited between the desired maximum (typically 5) and minimum (typically 2) values according to the following conditions:

D=D_{SNR}
if D < D_{min}

$D=D_{min}$

if $D > D_{max}$

$D=D_{max}$

After this a threshold adaptation coefficient ta_0 is calculated by

$$ta_0 = th_{max} - \frac{D-D_{min}}{D_{max}-D_{min}} (th_{max} - th_{min}), \quad (28)$$

where th_{min} and th_{max} are the minimum (typically 0.5) and maximum (typically 1) scaler values, respectively.

The actual scaler for frame n, $ta(n)$, is calculated by smoothing ta_0 with a filter with different time constants for increasing and decreasing values. The smoothing may be performed according to following equations:

if $ta_0 > ta(n\text{-}1)$

$$ta(n) = \lambda_0 ta(n\text{-}1)+(1-\lambda_0)ta_0 \quad (29)$$

else

$$ta(n)=\lambda_1 ta(n\text{-}1)+(1-\lambda_1)ta_0$$

Here λ_0 and λ_1 are the attack (increase period; typical value 0.9) and release (decrease period; typical value 0.5) time constants. Finally, the scaler $ta(n)$ can be used to scale the threshold vth in order to obtain a new VAD threshold value vth, whereby

$$vth = ta(n) \cdot vth2 \quad (30)$$

An often occurring problem in a voice activity detector is that just at the beginning of speech the speech is not detected immediately and also the end of speech is not detected correctly. This, on its behalf, causes that the background noise estimate $N(s)$ gets an incorrect value, which again affects later results of the voice activity detector. This problem can be eliminated by updating the background noise estimate using a delay. In this case a certain number N (e.g. N=2) of power spectra (here calculation spectra) $S_1(s),...,S_N(s)$ of the last frames are stored (e.g. in a buffer implemented at the input of block 80, not shown in figure 11) before updating the background noise estimate $N(s)$. If during the last double amount of frames (or during 2*N frames) the voice activity detector has not detected speech, the background noise estimate $N(s)$ is updated with the oldest power spectrum $S_1(s)$ in memory, in any other case updating is not done. With this it is ensured, that N frames before and after the frame used at updating have been noise.

The method according to the invention and the device for voice activity detection are particularly suitable to be used in communication devices such as a mobile station or a mobile communication system (e.g. in a base station), and they are not limited to any particular architecture (TDMA, CDMA, digital/analog). Figure 13 presents a mobile station according to the invention, in which voice activity detection according to the invention is employed. The speech signal to be transmitted, coming from a microphone 1, is sampled in an A/D converter 2, is speech coded in a speech codec 3, after which base frequency signal processing (e.g. channel encoding, interleaving), mixing and modulation into radio frequency and transmittance is performed in block TX. The voice activity detector 4 (VAD) can be used for controlling discontinous transmission by controlling block TX according to the output V_{ind} of the VAD. If the mobile station includes an echo and/or noise canceller ENC, the VAD 4 according to the invention can also be used in controlling block ENC. From block TX the signal is transmitted through a duplex filter DPLX and an antenna ANT. The known operations of a reception branch RX are carried out for speech received at reception, and it is repeated through loudspeaker 9. The VAD 4 could also be used for controlling any reception branch RX operations, e.g. in relation to echo cancellation.

Here realization and embodiments of the invention have been presented by examples on the method and the device. It is evident for a person skilled in the art that the invention is not limited to the details of the presented embodiments and that the invention can be realized also in another form without deviating from the characteristics of the invention. The presented embodiments should only be regarded as illustrating, not limiting. Thus the possibilities to realize and use the invention are limited only by the enclosed claims. Hereby different alternatives for the implementing of the invention defined by the claims, including equivalent realizations, are included in the scope of the invention.

Claims

1. A voice activity detection device comprising

means for detecting voice activity in an input signal (x(n)), and
means for making a voice activity decision (V_{ind}) on basis of the detection, **characterized** in that it comprises

means (6) for dividing said input signal (x(n)) in subsignals (S(s)) representing specific frequency bands,
means (80) for estimating noise (N(s)) in the subsignals,

means (90) for calculating subdecision signals (SNR(s)) on basis of the noise in the subsignals, and
means (110) for making a voice activity decision (V_{ind}) for the input signal on basis of the subdecision signals.

2. A voice activity detection device according to claim 1, **characterized** in that it comprises means (90) for calculating a signal-to-noise ratio (SNR) for each subsignal and for providing said signal-to-noise ratios as subdecision signals (SNR(s)).

3. A voice activity detection device according to claim 2, **characterized** in that the means (110) for making a voice activity decision (V_{ind}) for the input signal comprises

means (111) for creating a value (D_{SNR}) based on said signal-to-noise ratios (SNR(s)), and
means (112) for comparing said value (D_{SNR}) with a threshold value (vth) and for outputting a voice activity decision signal (V_{ind}) on basis of said comparison.

4. A voice activity detection device according to claim 1, **characterized** in that it comprises means (70) for determining the mean level of a noise component and a speech component ($\hat{N},\hat{S}$) contained in the input signal, and means (113) for adjusting said threshold value (vth) based upon the mean level of the noise component and the speech component ($\hat{N},\hat{S}$).

5. A voice activity detection device according to claim 2, **characterized** in that it comprises means (113) for adjusting said threshold value (vth) based upon past signal-to-noise ratios (SNR(s)).

6. A voice activity detection device according to claim 2, **characterized** in that it comprises means (80) for storing the value of the estimated noise (N(s)) and said noise (N(s)) is updated with past subsignals (S(s)) depending on past and present signal-to-noise ratios (SNR(s)).

7. A voice activity detection device according to claim 1, **characterized** in that it comprises means (3) for calculating linear prediction coefficients based on the input signal (x(n)), and means (8) for calculating said subsignals (S(s)) based on said linear prediction coefficients.

8. A voice activity detection device according to claim 1, **characterized** in that it comprises

means (7) for calculating a long term prediction analysis producing long term predictor parameters, said parameters including long term predictor gain (LTP_gain_sum),
means (7) for comparing said long term predictor gain with a threshold value (thr_lag), and
means for producing a voiced detection decision on basis of said comparison.

9. A mobile station for transmission and reception of speech messages, comprising

means for detecting voice activity in a speech message (x(n)), and
means for making a voice activity decision (V_{ind}) on basis of the detection, **characterized** in that it comprises

means (6) for dividing said speech message (x(n)) in subsignals (S(s)) representing specific frequency bands,
means (80) for estimating noise (N(s)) in the subsignals,
means (90) for calculating subdecision signals (SNR(s)) on basis of the noise in the subsignals, and
means (110) for making a voice activity decision (V_{ind}) for the input signal on basis of the subdecision signals.

10. A method of detecting voice activity in a communication device, the method comprising the steps of:

receiving an input signal (x(n)),
detecting voice activity in the input signal, and
making (110) a voice activity decision (V_{ind}) on basis of the detection, **characterized** in that it comprises
dividing (6) said input signal in subsignals (S(s)) representing specific frequency bands,
estimating noise (N(s)) in the subsignals,
calculating (90) subdecision signals (SNR(s)) on basis of the noise in the subsignals, and
making (110) a voice activity decision (V_{ind}) for the input signal on basis of the subdecision signals.

Fig.1

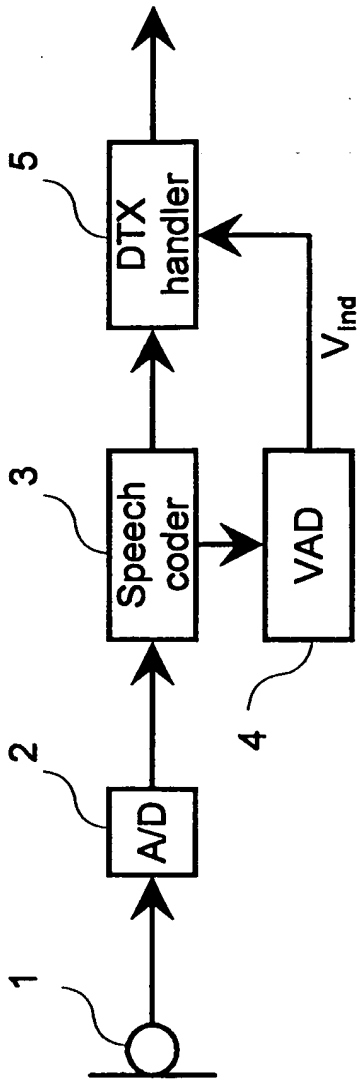

1
A/D
2
Speech coder
3
VAD
4
DTX handler
5
V_{ind}

Fig.5

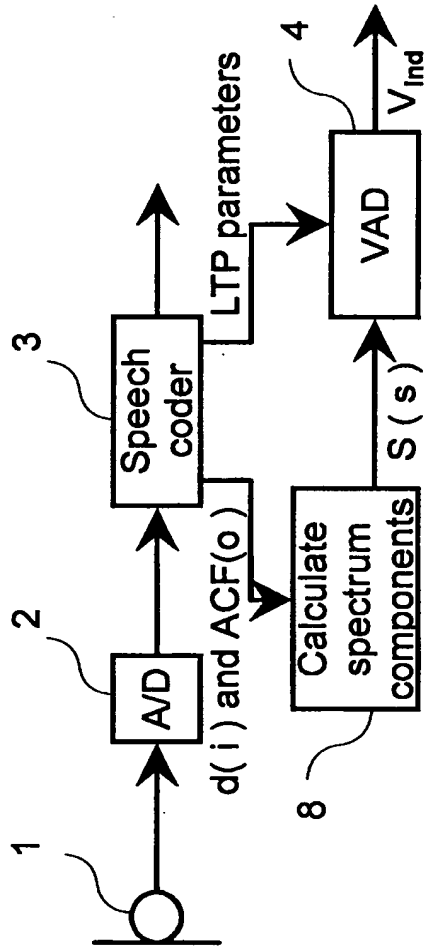

1
A/D
2
Speech coder
3
LTP parameters
VAD
4
V_{ind}
d(i) and ACF(o)
Calculate spectrum components
8
S (s)

Fig.12

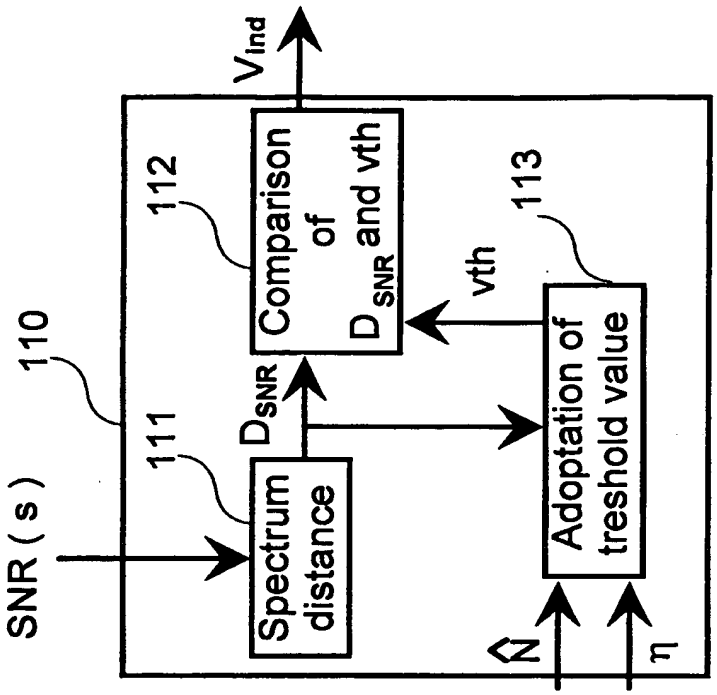

V_{ind}
110
112
Comparison of D_{SNR} and vth
113
vth
D_{SNR}
111
SNR (s)
Spectrum distance
Adoptation of treshold value
$\hat{N}$
η

Fig.2

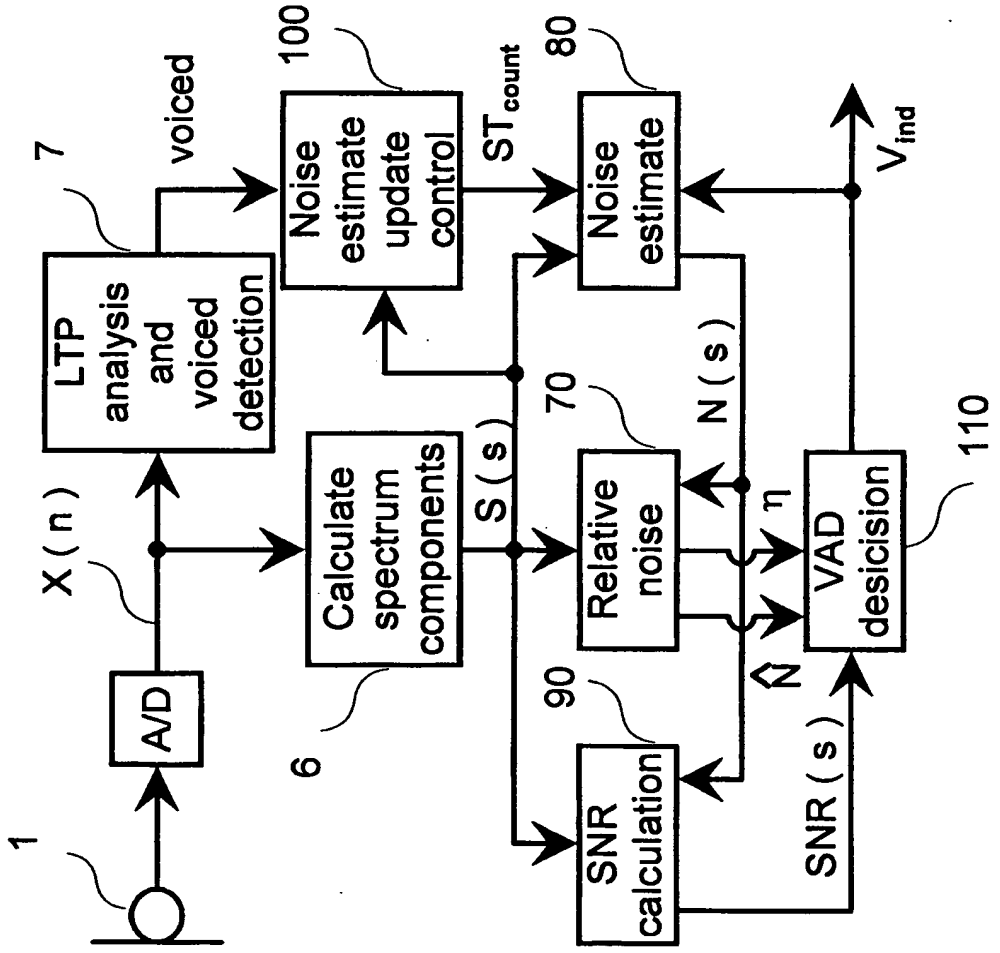
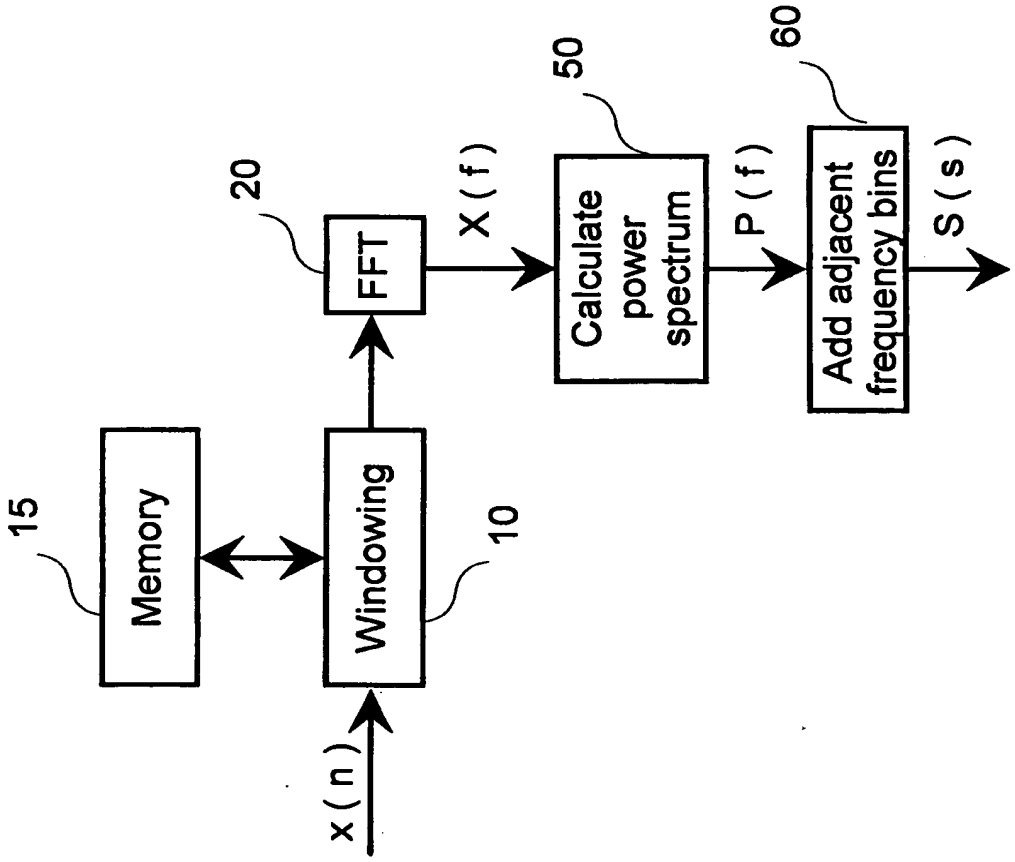

1
A/D
X (n)
LTP analysis and voiced detection
7
voiced
Calculate spectrum components
6
S (s)
Noise estimate update control
100
ST_{count}
Noise estimate
80
V_{ind}
Relative noise
70
N (s)
η
$\hat{N}$
VAD desicision
110
SNR calculation
90
SNR (s)

Fig.3


x (n)
Windowing
10
Memory
15
FFT
20
X (f)
Calculate power spectrum
50
P (f)
Add adjacent frequency bins
60
S (s)

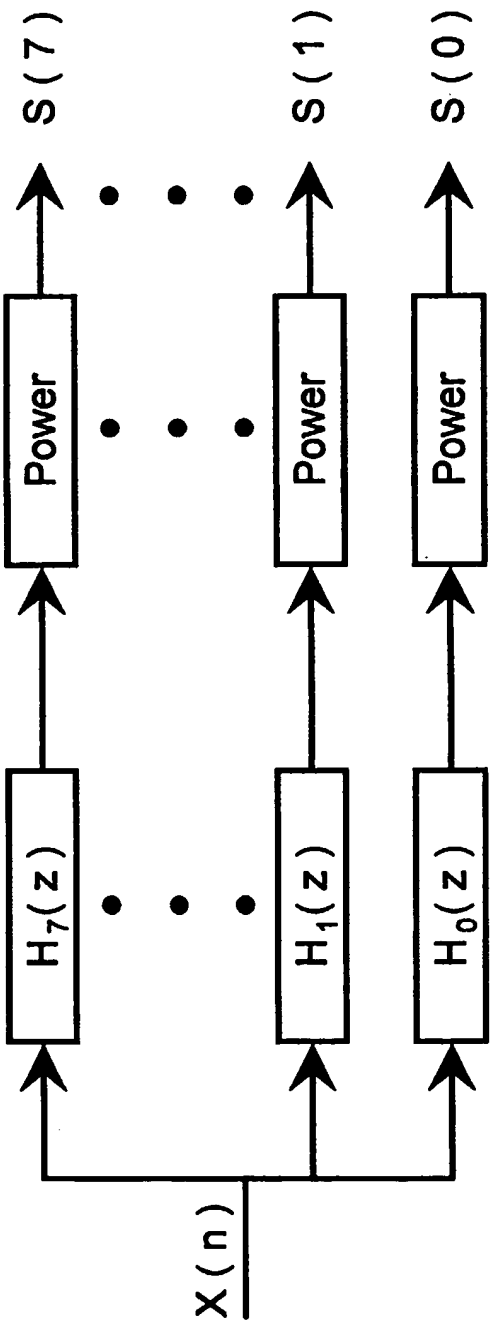
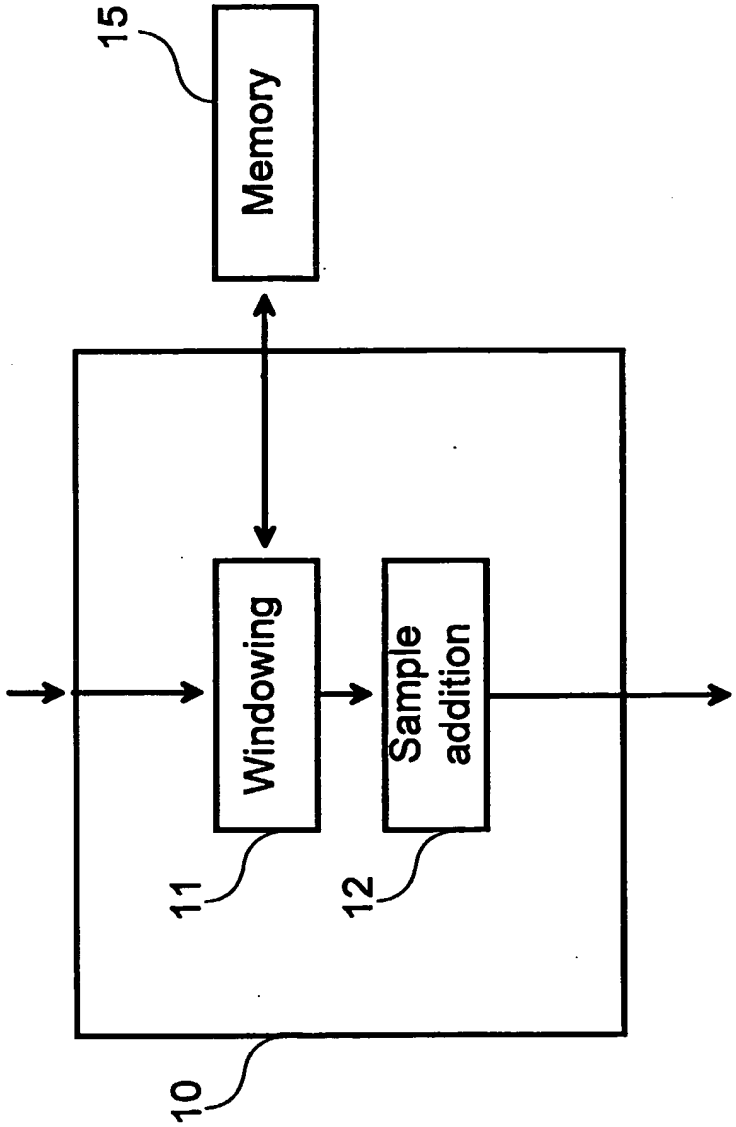


Fig.4

Fig. 6



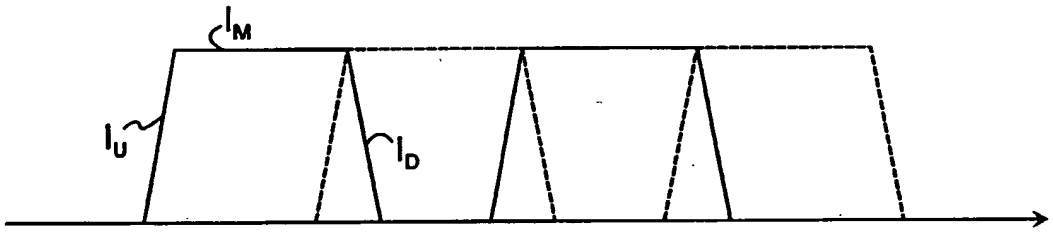


Fig. 7

Fig. 8

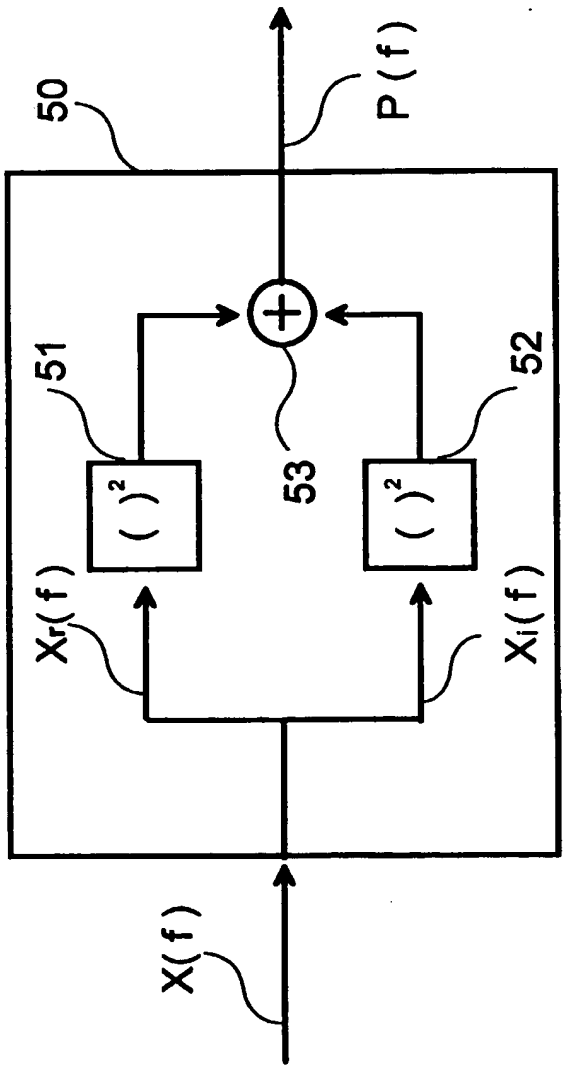
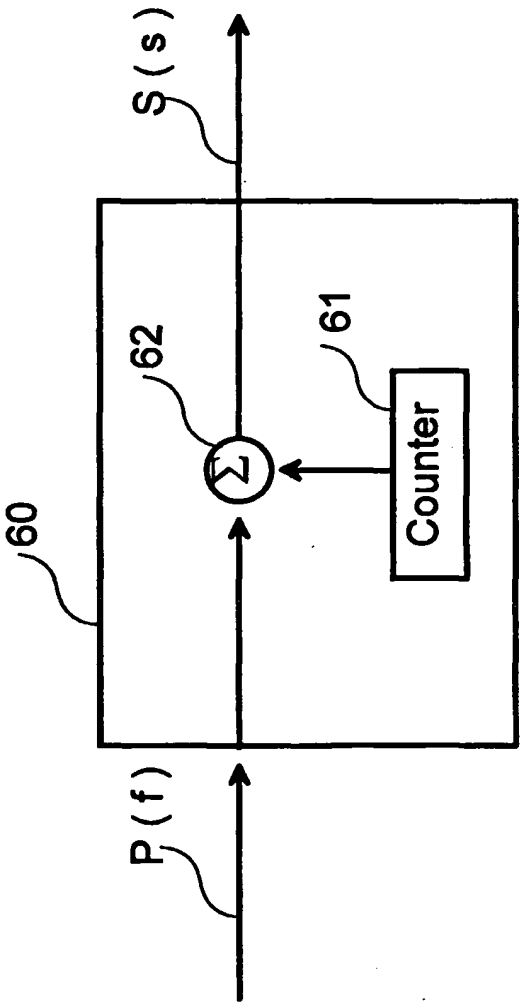


Fig. 9



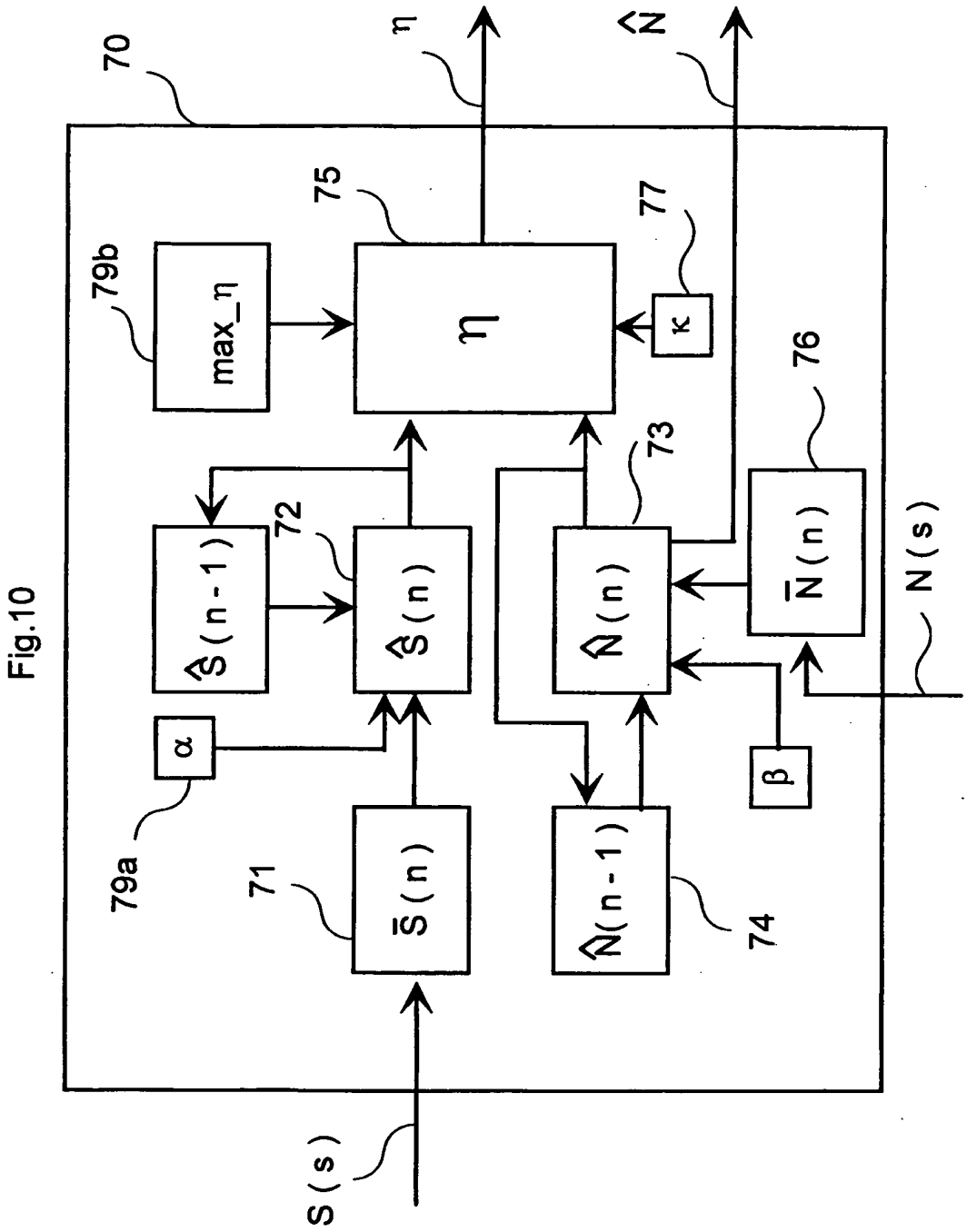
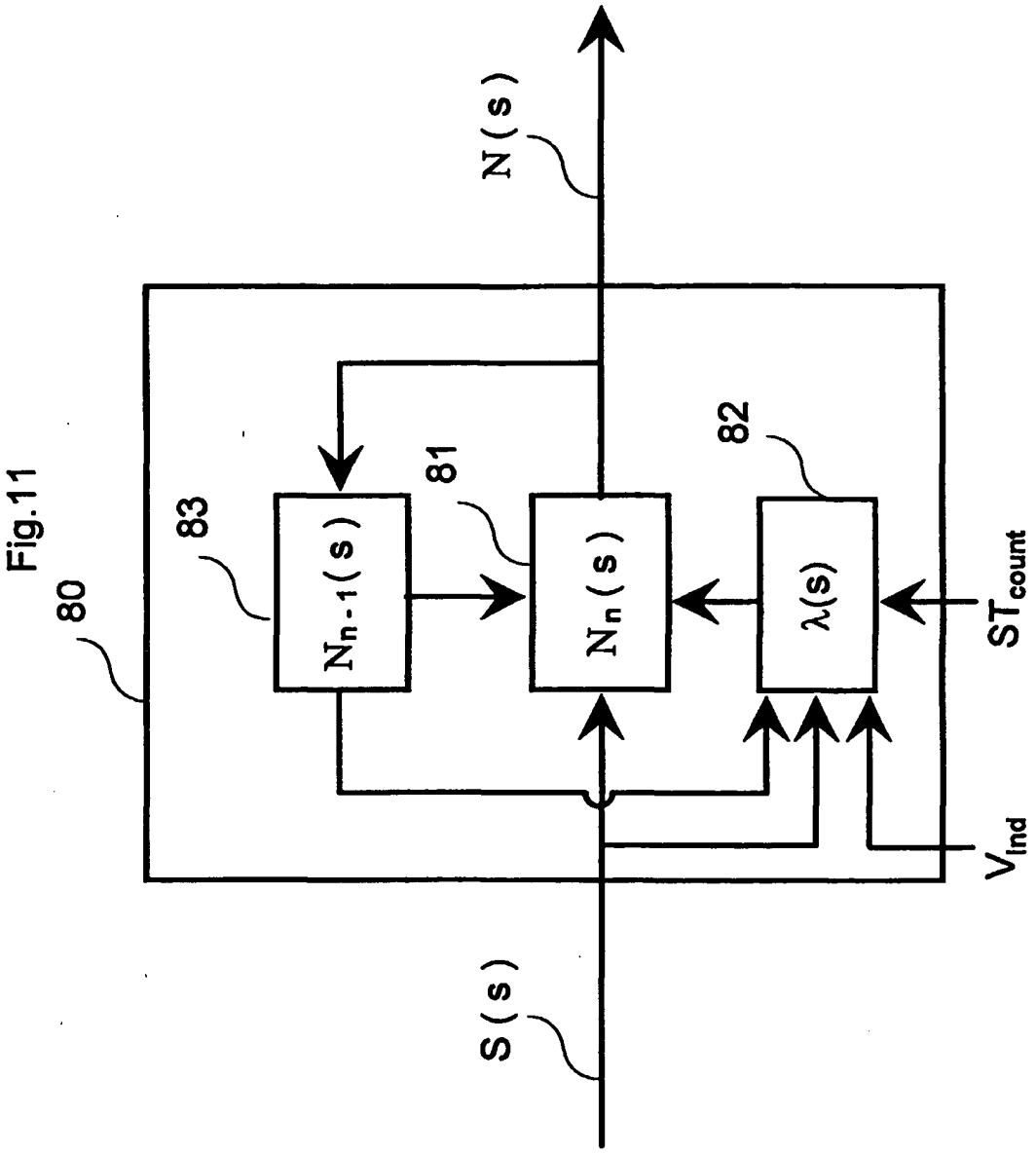

70
η
$\hat{N}$
75
77
79b
max_η
η
κ
76
73
72
$\hat{S}(n-1)$
$\hat{S}(n)$
$\hat{N}(n)$
$\bar{N}(n)$
N(s)
α
β
79a
71
74
$\bar{S}(n)$
$\hat{N}(n-1)$
S(s)

Fig.10

Fig.11



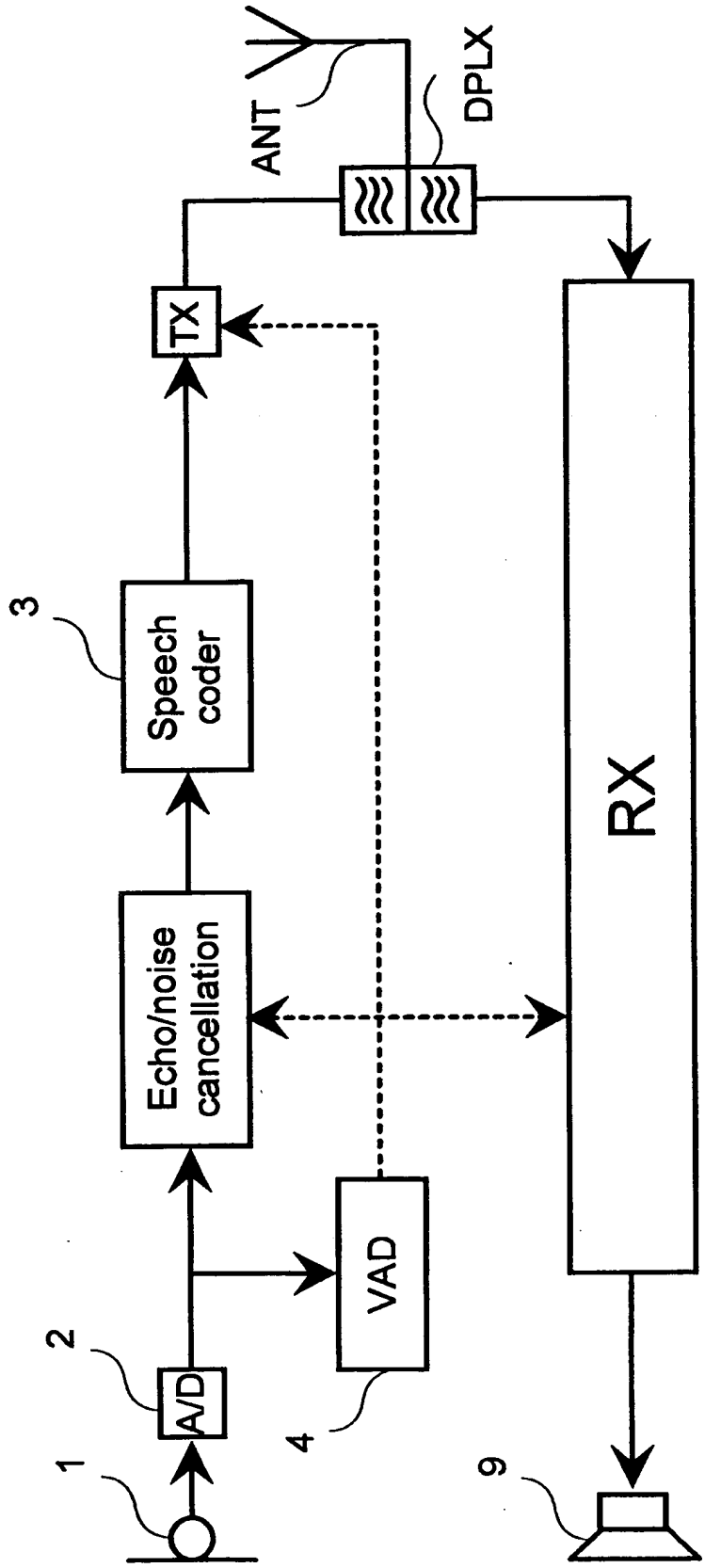

ANT
DPLX
TX
3
Speech coder
Echo/noise cancellation
RX
VAD
2
A/D
4
1
9

Fig.13

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 96 11 8504.
Page 1

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.6)
A	US 5459814 A (PRABHAT K. GUPTA ET AL), 17 October 1995 (17.10.95) * column 5, line 5 - column 6, line 48, figure 2 * --	1-10	G10L 3/00
D,A	US 5276765 A (DANIEL K. FREEMAN ET AL), 4 January 1994 (04.01.94) * column 3, line 49 - column 8, line 23, cited in the application * --	1-10	
A	EP 0222083 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION), 20 May 1987 (20.05.87) * column 1, line 33 - column 2, line 7; column 3, line 41 - column 8, line 49, figure 3 * -----	1-10	

TECHNICAL FIELDS SEARCHED (Int. Cl.6)
G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
STOCKHOLM	13 May 1997	CHRISTERSON LARS

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P0401)